



Agriculture and
Agri-Food Canada

Agriculture et
Agroalimentaire Canada



Tools and Approaches for SNP discovery

Wayne Clarke

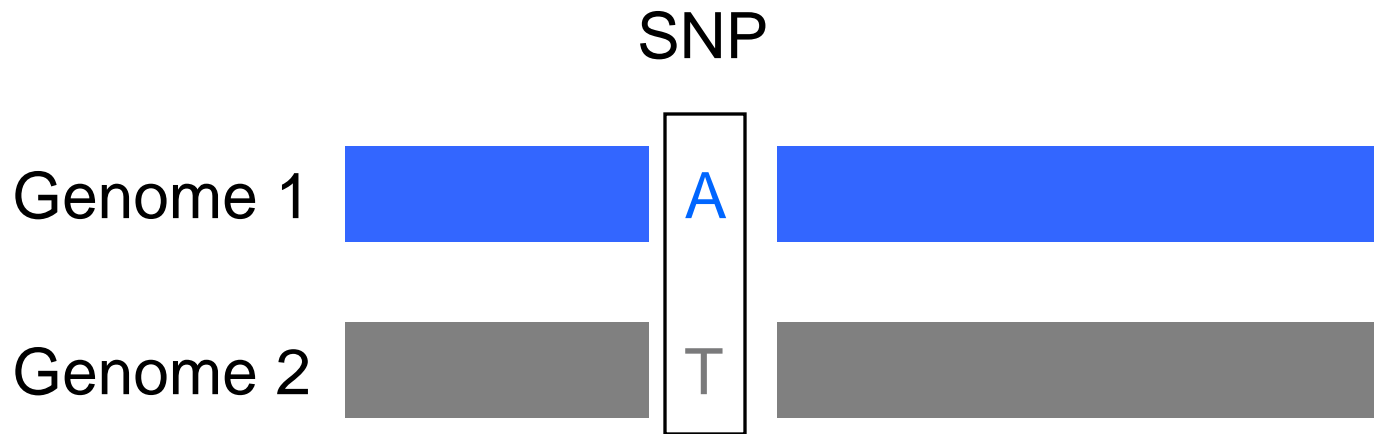
Agriculture and Agri-Food Canada, Research Centre, Saskatoon, SK
Ph.D. Candidate, Department of Computer Science,
University of Saskatchewan



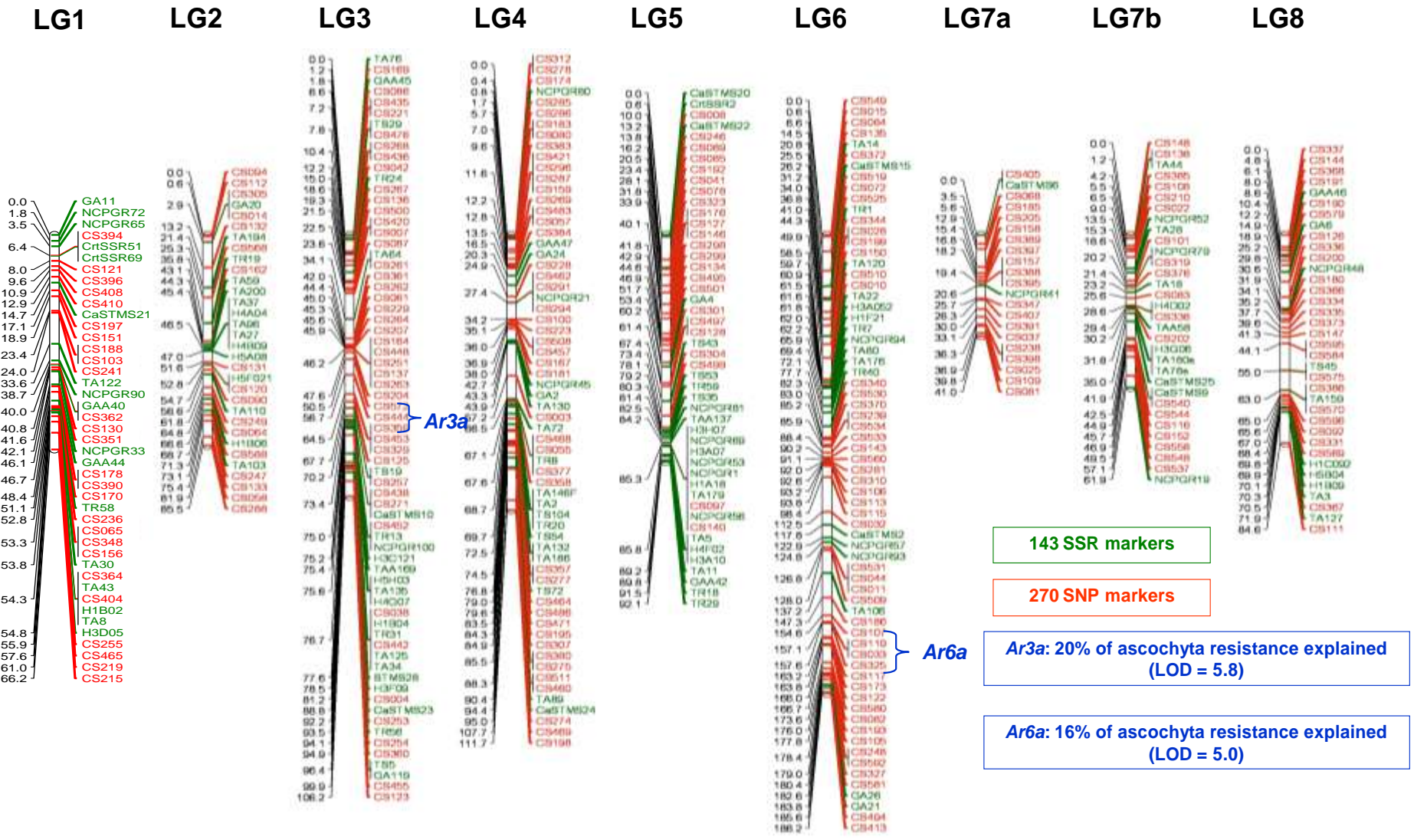
Outline

- Introduction
- 3'Capture Sequencing Technique
- Lentil SNP discovery
- Brassica SNP discovery
- Future Work
- Challenges of SNP discovery

Single Nucleotide Polymorphisms (SNPs)



Genetic Mapping with SNPs



143 SSR markers

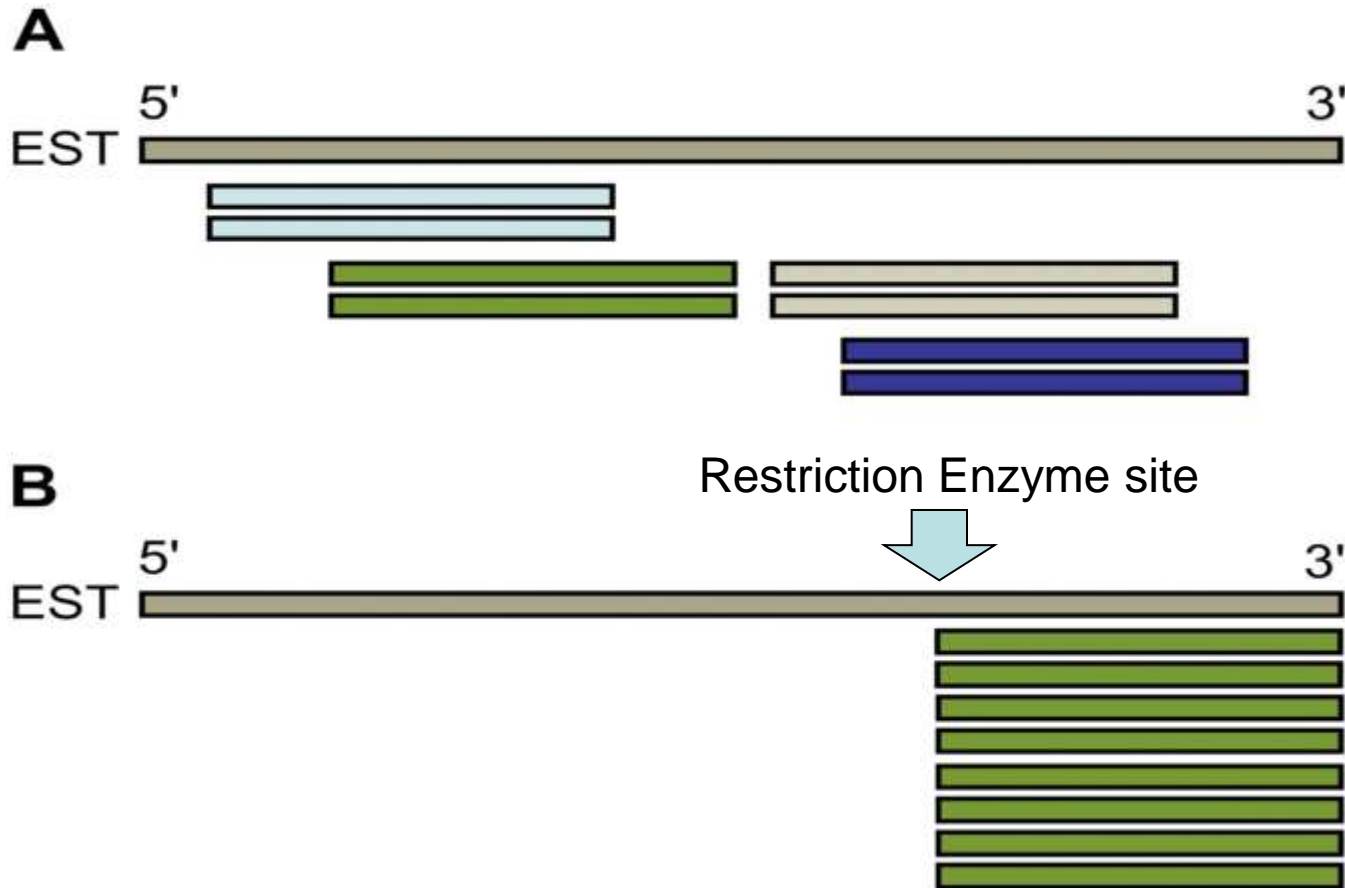
270 SNP markers

Ar3a: 20% of ascochyta resistance explained (LOD = 5.8)

Ar6a: 16% of ascochyta resistance explained (LOD = 5.0)

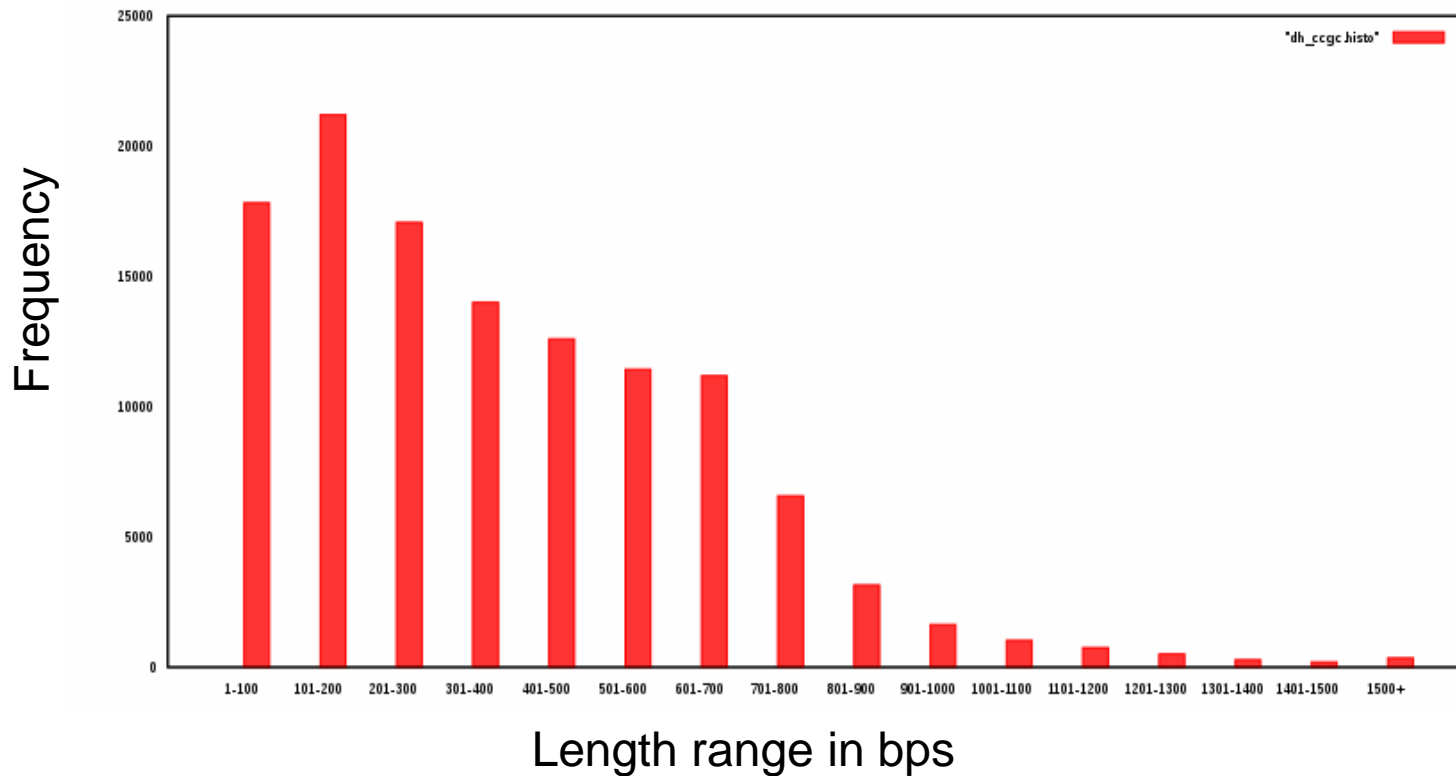
3'Capture Sequencing Technique

Shotgun Sequencing vs. Anchored Approach



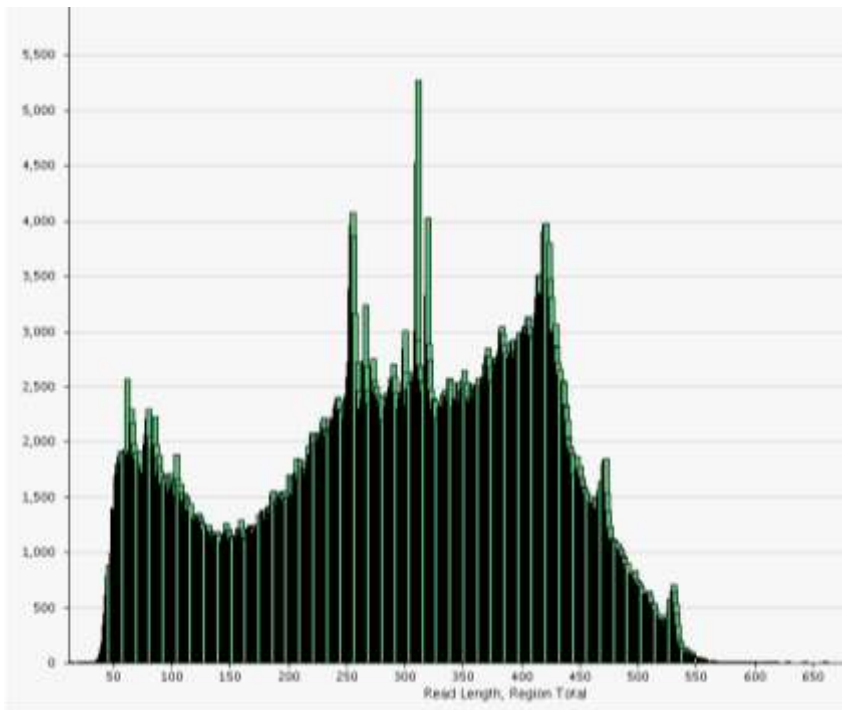
3'Capture Sequencing Technique

- Frequency of fragment sizes after *in silico* digestion of *B. napus* ESTs using *Aci* I



3'Capture Sequencing Technique

- Sample output from 3' sequencing run



TCAG (Library)	Region		Total
	1	2	
Raw Wells	925,128	819,690	1,744,818
Key Pass Wells	901,755	800,841	1,702,596
Passed Filter Wells	472,899	483,140	956,039
Total Bases	139,787,293	138,732,243	278,519,536
Length Average	295.60	287.15	291.33
Length Std Deviation	129.06	119.99	
Longest Reads Length	617	761	761
Shortest Reads Length	21	12	12
Median Reads Length	314.0	295.0	304.0

Experimental Procedure

- Generation of 3'capture data
- Determine high quality SNPs
- Validation of SNPs using the KASPar system
- Array Design

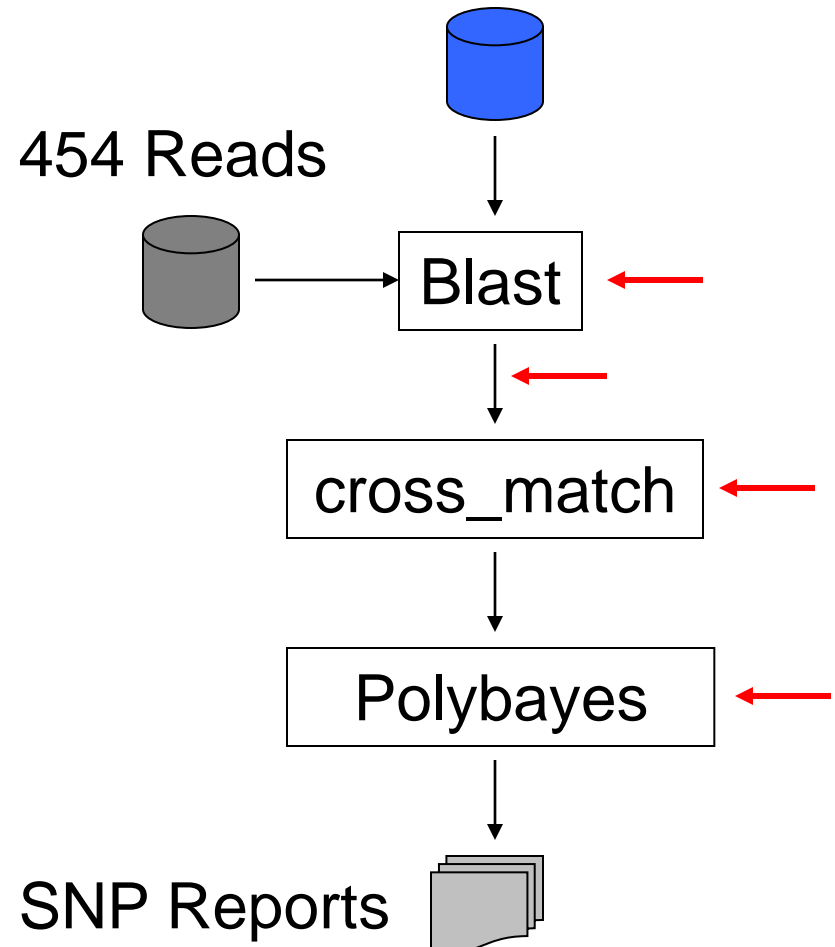
Lentil SNP Discovery

- 3'Capture was performed for 11 genotypes
- Chose a single line (CDC_Redberry) as the reference and scored SNPs in the remaining genotypes against this reference
- Tried 3 different SNP calling methods

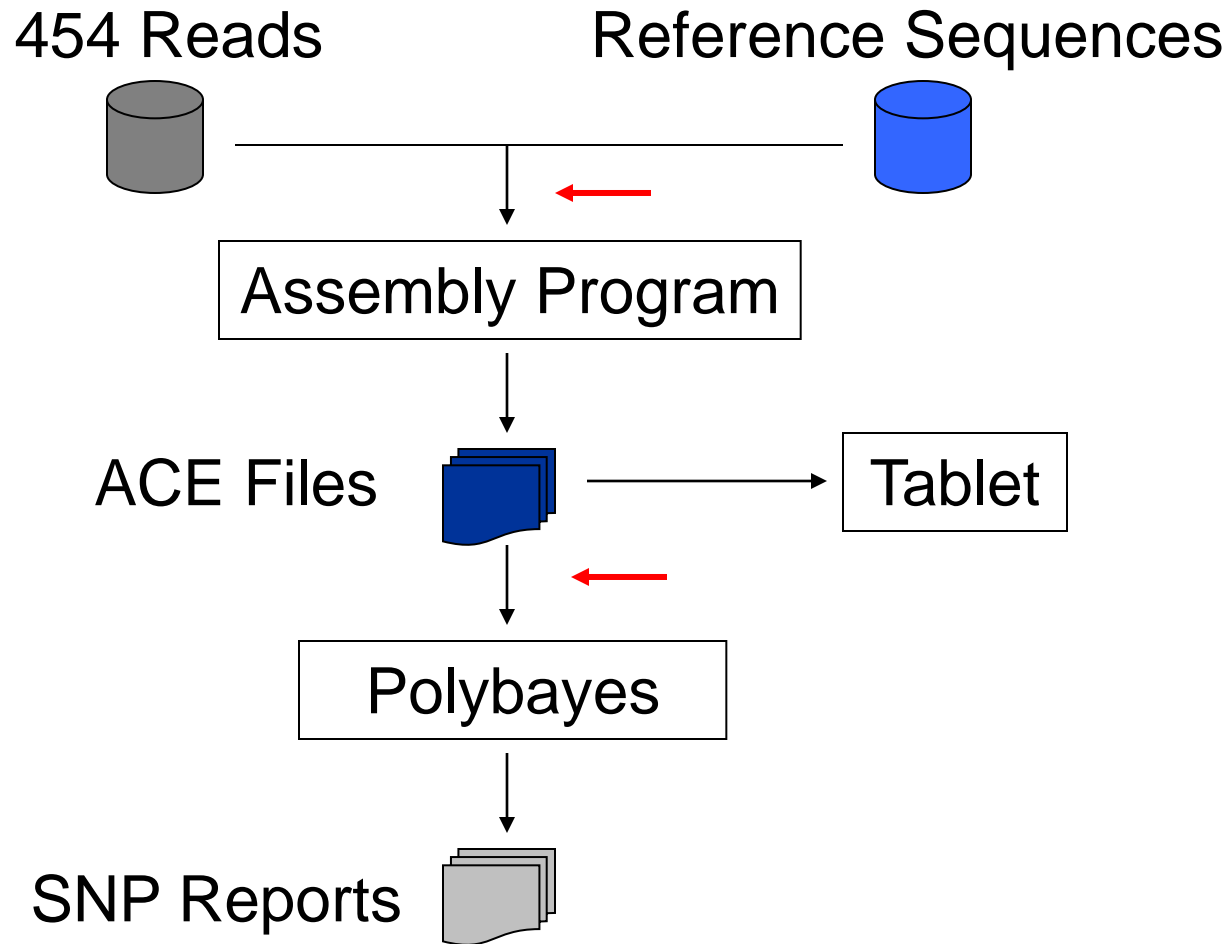
Barbazuk et al.

- Open Source
 - BLAST
 - cross_match
 - Polybayes
- **Pros**
 - Freely Available tools
 - Moderate running time
- **Cons**
 - Requires more user intervention than other methods
 - No good method for visualizing alignment information to confirm SNPs

Reference Sequence Database



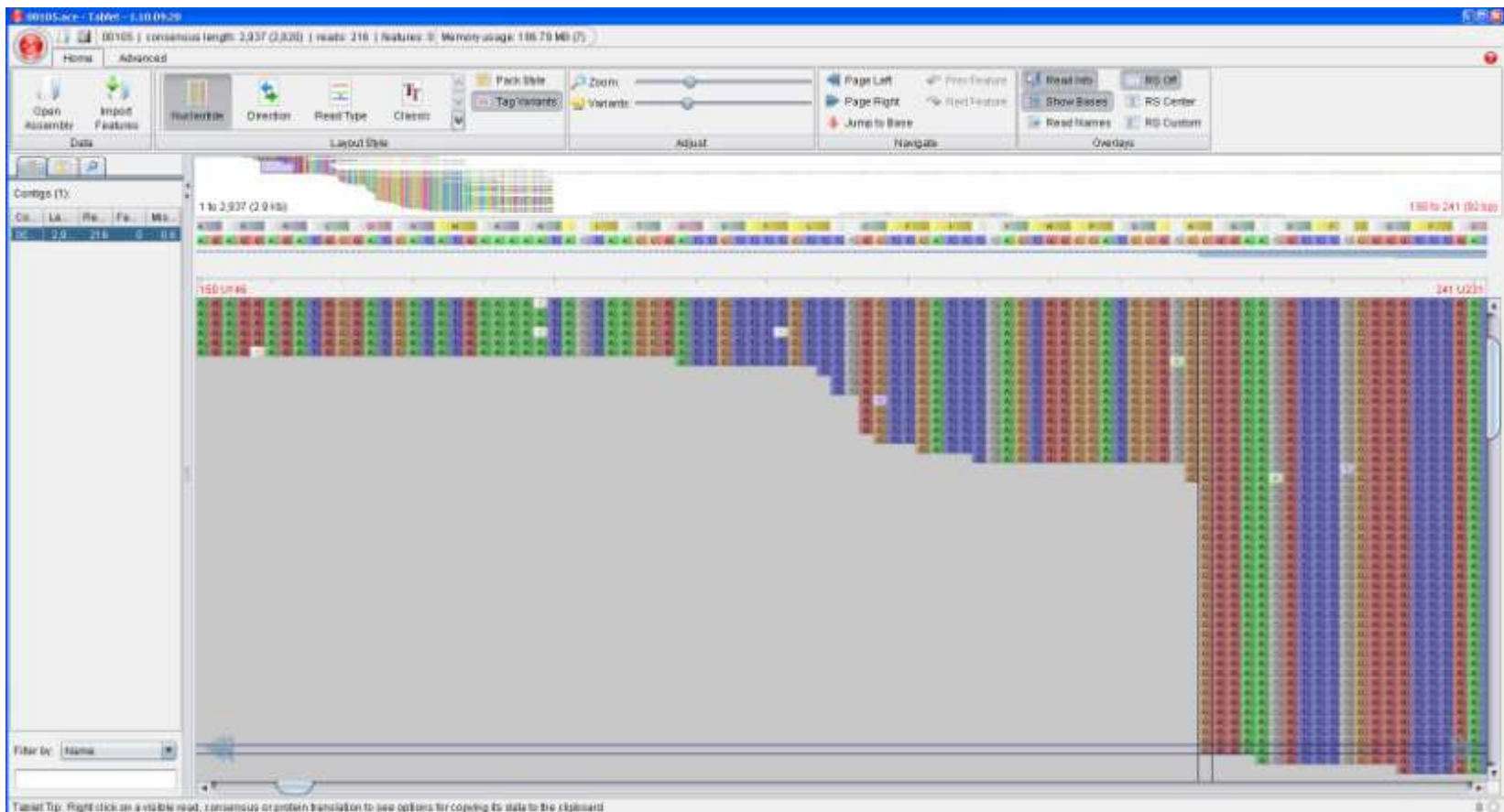
Ace + Polybayes



Tablet (<http://bioinf.scri.ac.uk/tablet/>)

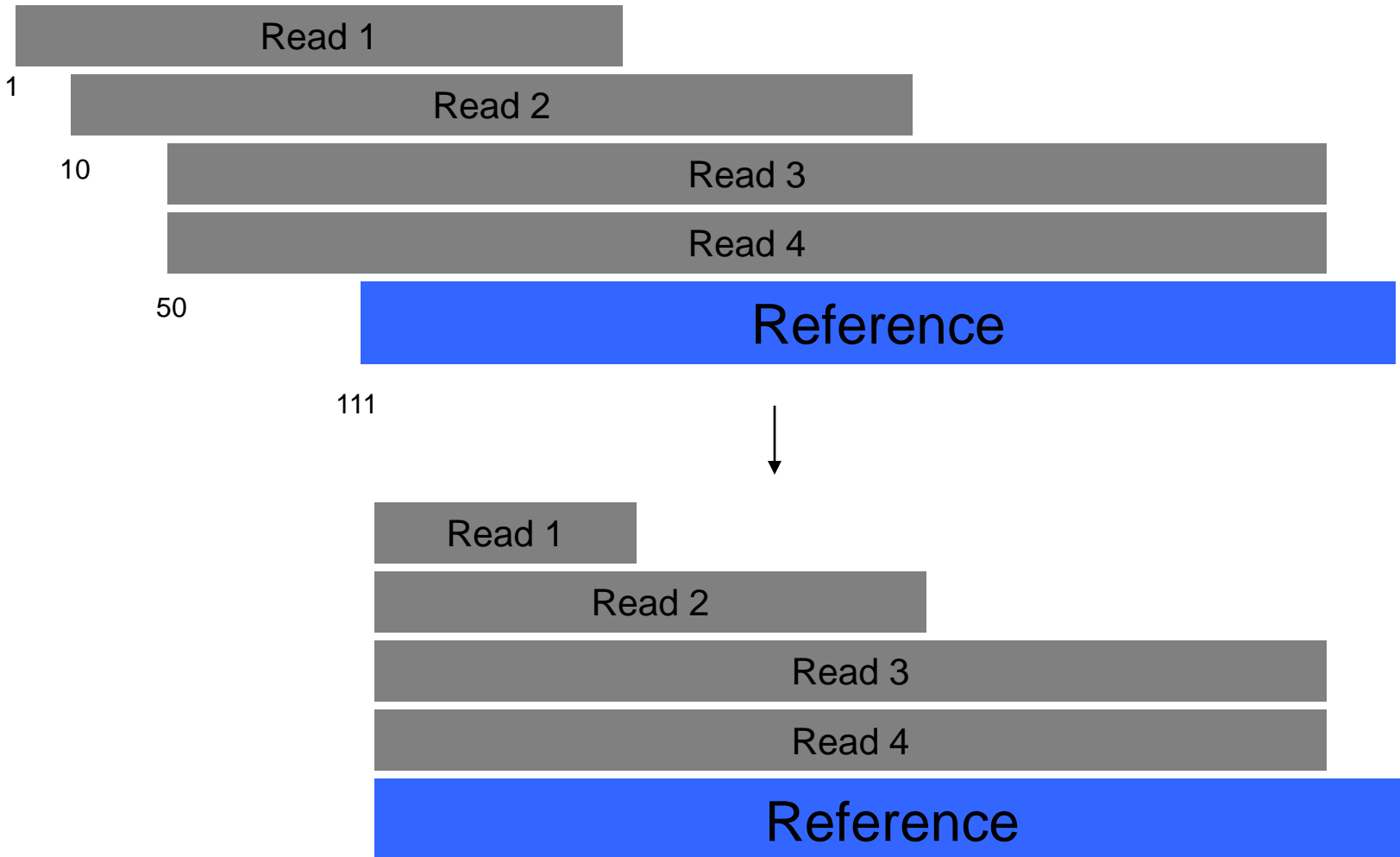
Ace + Polybayes

Problem: Some assemblers produces an alignment that may be extended past the end of the reference sequence



ACE + Polybayes

Solution: Trim the ends of all contigs in the ACE files



ACE + Polybayes

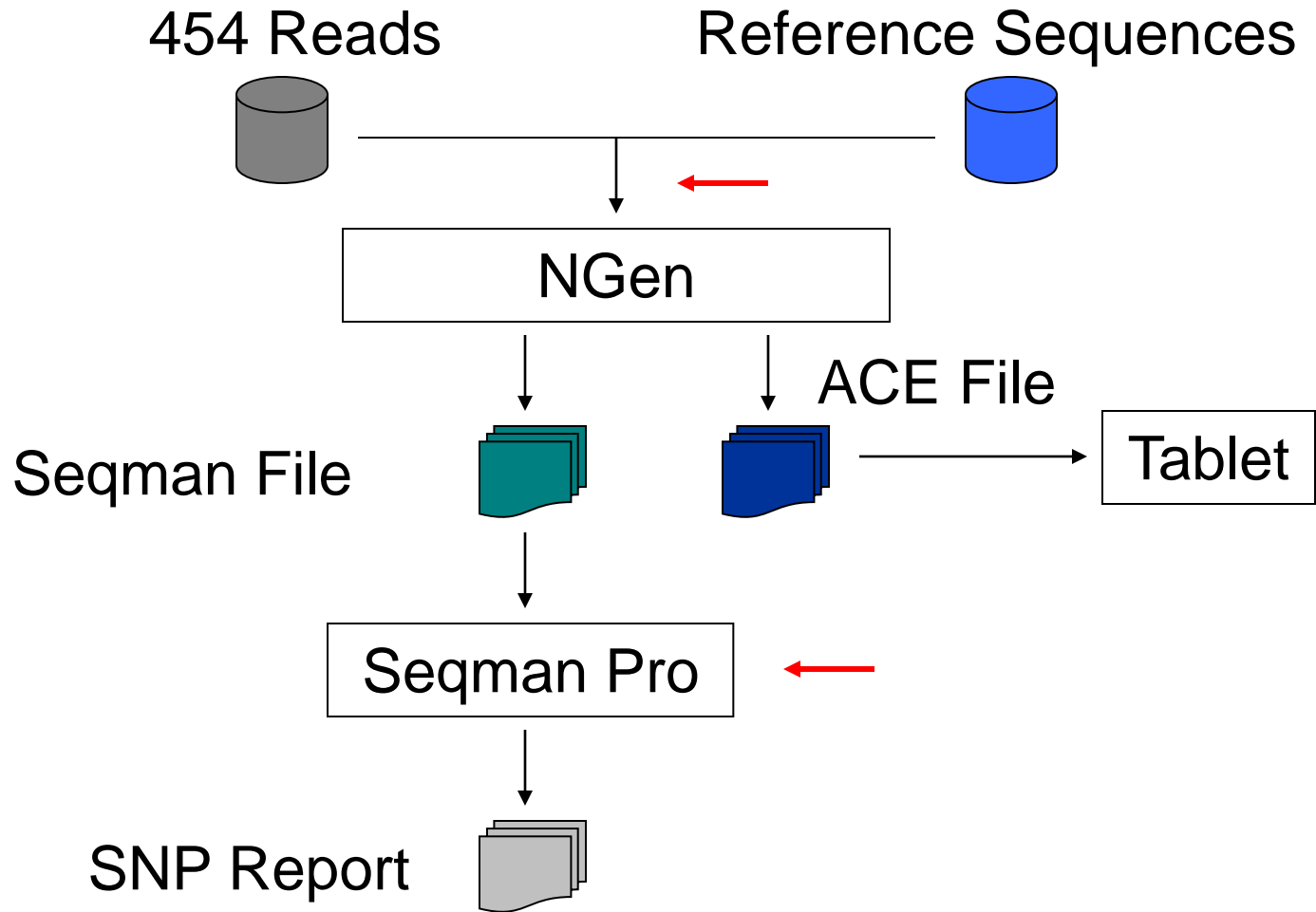
- **Pros**

- Do not have to buy any software
- Easy to visualize the data to manually confirm SNPs

- **Cons**

- Trimming contigs in ACE files is very time consuming
- Trimming adds more manual intervention to the process
- Have experienced some inconsistent results

DNASar



Seqman NGen (<http://www.dnastar.com/t-products-seqman-ngen.aspx>)

Seqman Pro (<http://www.dnastar.com/t-sub-products-lasergene-seqmanpro.aspx>)

DNAStar

- NGen Assembler

964a-46_vs_ref.script - NGen

File Edit Help

Workflow Reads Controls Actions Script

Please provide some information about your project before adding your sequence files.

Assembly Output

Where would you like to save the assembly?

Project name (provides a name for assembly, report, and other files)

Project folder Browse...

Type of Assembly

Template / Reference

De novo Genome length bp

De novo Transcriptome

Iterative assembly

Open SeqMan/ACE project Browse... Clear

Import SOLID GFF file

Read Technology: (sets default values for parameters on the next pages)

Actions

Do quality trimming

Do vector/adaptor scan

See the next page to add sequence files.

When the color changes from red to green, the script is ready to run.

Help < Previous Next > Assemble

Welcome to NGen!

964a-46_vs_ref.script - NGen

File Edit Help

Workflow Reads Controls Actions Script

Parameter values are based on a template/reference workflow.

Parameters Scans

Basic Advanced

Repeat Handling

Use repeat handling

Expected coverage

Genome length bp

Match Quality

High match percent

Normal match percent

Low match percent

Specific match threshold %

Match Size

High match size

Normal match size

Specific match size

When the color changes from red to green, the script is ready to run.

Help < Previous Next > Assemble

Welcome to NGen!

DNAStar

- Seqman Pro

The screenshot displays the SeqMan Pro software interface. The main window is titled "964a-46_vs_ref.sq". The menu bar includes "File", "Edit", "Sequence", "Contig", "Project", "Features", "SNP", "Net Search", "Window", and "Help". The "SNP" menu is open, showing options: "SNP Report", "Show SNPs", "Sort by SNP", "Confirm SNP", "Reject SNP", and "Putative SNP".

The main window displays a table of SNP analysis results. The table has columns for "Name", "Length", "Seqs", "Pos", and "Conflict Split". The data is organized into a section titled "Unlocated Contigs".

Name	Length	Seqs	Pos	Conflict Split
Unlocated Contigs				
<input type="checkbox"/> 00001	468	13	0	?
<input type="checkbox"/> 00002	613	111	0	?
<input type="checkbox"/> 00003	712	19	0	?
<input type="checkbox"/> 00004	479	23	0	?
<input type="checkbox"/> 00005	1340	45	0	?
<input type="checkbox"/> 00006	1134	57	0	?
<input type="checkbox"/> 00007	612	42	0	?
<input type="checkbox"/> 00008	464	9	0	?
<input type="checkbox"/> 00009	1062	35	0	?
<input type="checkbox"/> 00010	674	55	0	?

Below the table, a list of contig IDs is displayed:

- 00001
- 00002
- 00003
- 00004
- 00005
- 00006
- 00007
- 00008
- 00009
- 00010
- 00011
- 00012
- 00013

DNAStar

- Seqman Pro

SeqMan

File Edit Sequence Contig Project Features SNP Net Search Window Help

964a-46_vs_ref.sqg

SNP Statistics Report from Contigs

All Found SNPs | SNPs Summary

Confirmed SNP
 Putative SNP
 Rejected SNP
 Mixed SNP

Show All SNPs

Show Counts as a percent

SNP Percent Filter keep range min. to max. Depth Max. Coding Feature Distance

237456 SNP Columns Rejected: 0 Confirm: 0 Putative: 237456 Mixed: 0 Filtered: 0

SNP	Contig ID	Contig Pos	Ref Pos	Type	Ref Base	SNP Base	SNP %	DBSNP ID	Codon	ture Na
?	22190	5	5	Indel	A	-	33.3 %			
?	22190	81	81	Indel	A	-	33.3 %			
?	22190	82	82	Indel	A	-	66.7 %			
?	22190	122	122	Indel	A	-	50.0 %			
?	22190	149	149	Indel	-	A	50.0 %			
?	22190	150	149	Indel	-	T	50.0 %			
?	22190	179	177	Indel	-	T	100.0 %			
?	22998	17	10	SNP	G	A	6.3 %			
?	22998	18	11	Indel	T	-	50.0 %			
?	22998	21	14	SNP	T	A	5.9 %			
?	22998	22	15	SNP	T	G	5.9 %			
?	22998	24	17	Indel	T	-	5.9 %			
?	22998	25	18	Indel	T	-	17.6 %			
?	22998	26	19	Indel	T	-	52.9 %			
?	22998	27	20	Indel	T	-	70.6 %			
?	22998	32	25	Indel	-	T	17.6 %			
?	22998	64	56	Indel	A	-	3.0 %			
?	22998	67	59	Indel	C	-	3.0 %			
?	22998	73	65	Indel	-	G	3.1 %			
?	22998	76	67	SNP	G	C	3.1 %			
?	22998	77	68	Indel	G	-	6.3 %			
?	22998	79	70	Indel	-	T	9.4 %			
?	22998	82	72	Indel	T	-	3.1 %			
?	22998	99	89	Indel	-	T	13.3 %			
?	22998	111	100	Indel	-	T	3.4 %			

DNAStar

- **Pros**
 - Ease of use
 - NGen
- **Cons**
 - Proprietary Software
 - Seqman Pro
 - Reference extension

Brassica napus

- 3'Capture data for 2 parental lines and 18 Segregation lines
- **High Quality reference contigs generated**
- Tried 3 methods for SNP calling

Complex Genomes

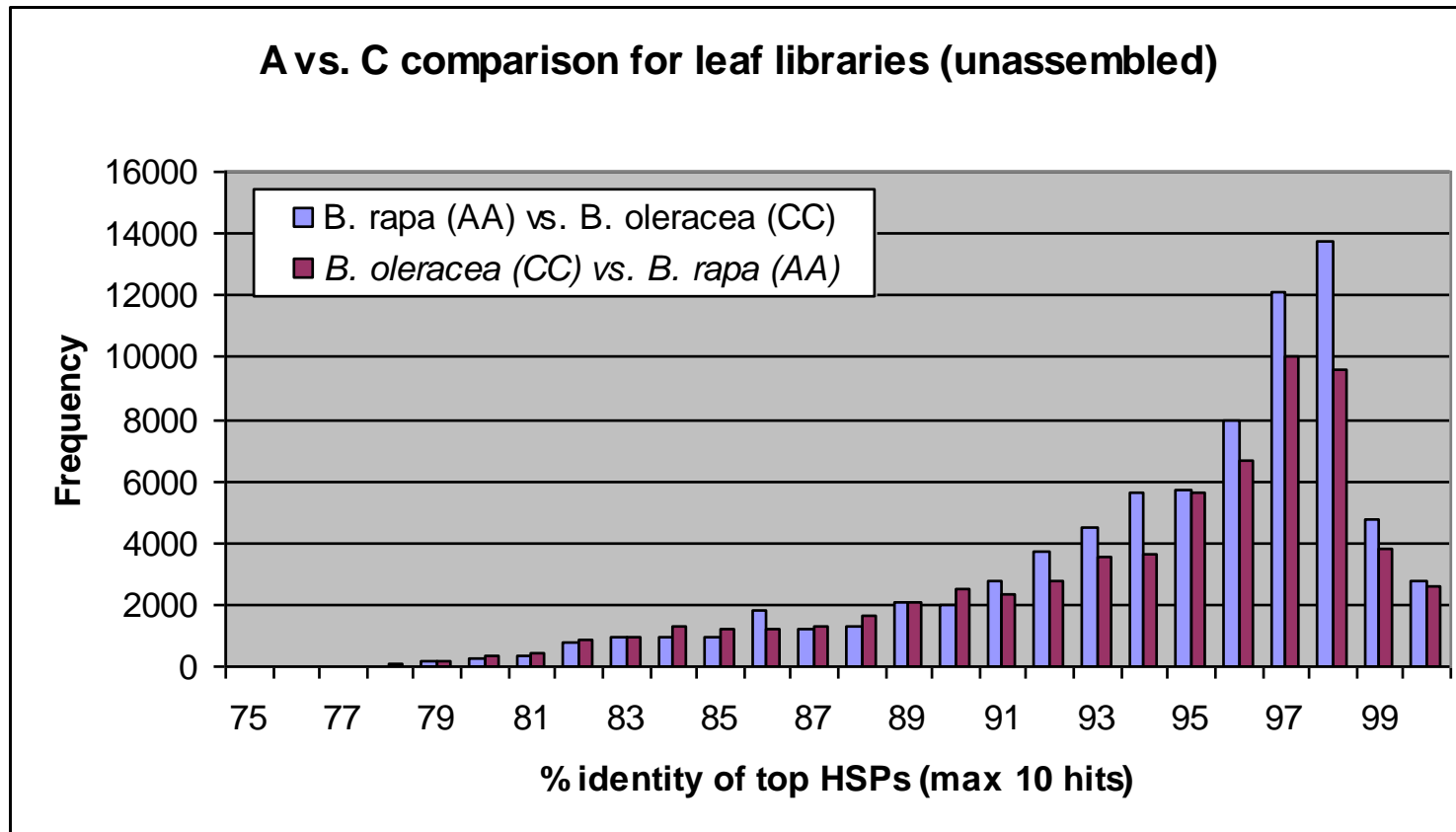
Solution: Increase stringency of the assembly of the reference as well as the stringency of the reference mapping

Optimizing assembly parameters to separate homeologues

- Use available *B. oleracea* and *B.rapa* ESTs
- Blast *oleracea* vs *rapa* then *rapa* vs *oleracea*
- Record the sequence similarity of the top alignments
- H0: the sequence similarity of the top alignments should be reflective of the ortholog similarity between the species

Complex Genomes

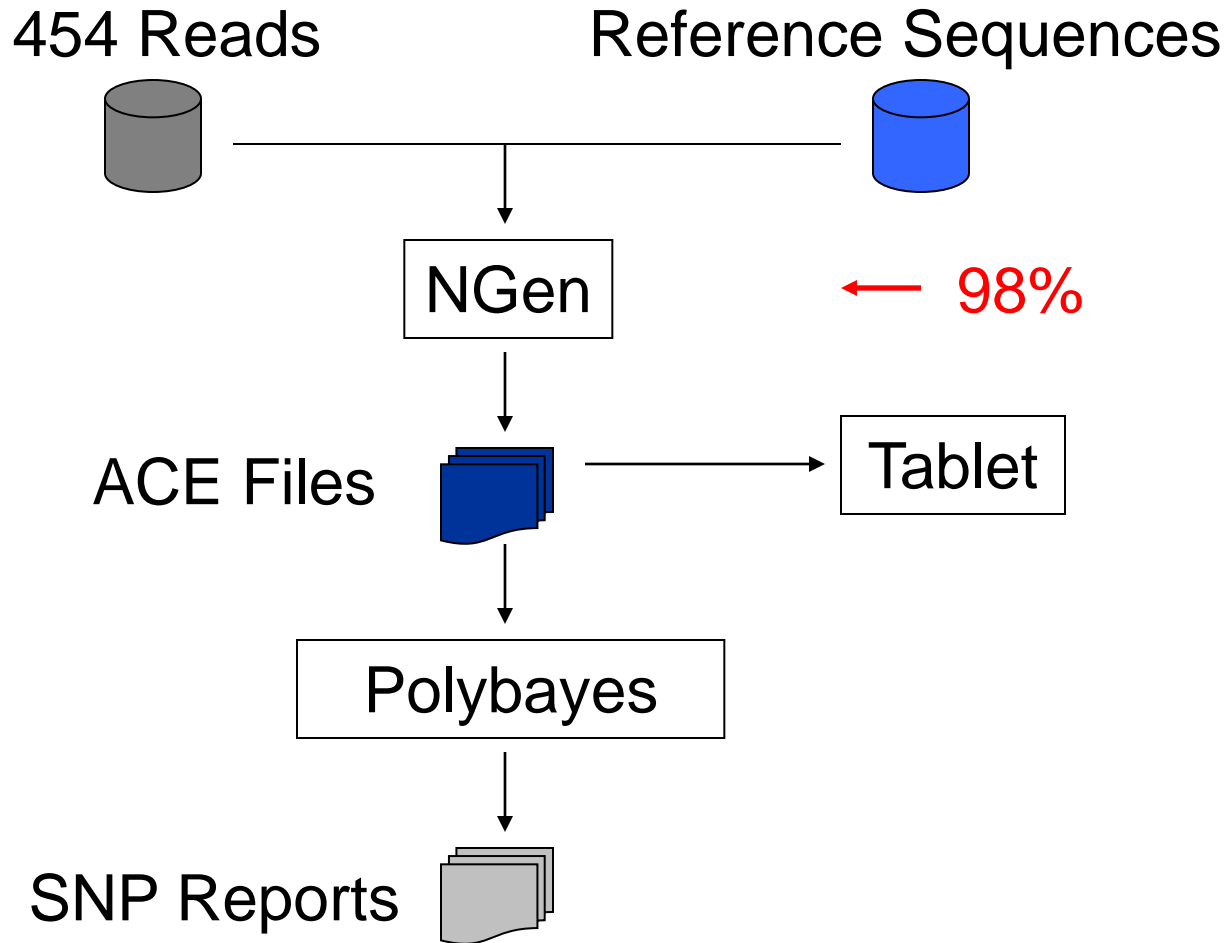
Optimizing assembly parameters to separate homeologues



Brassica napus

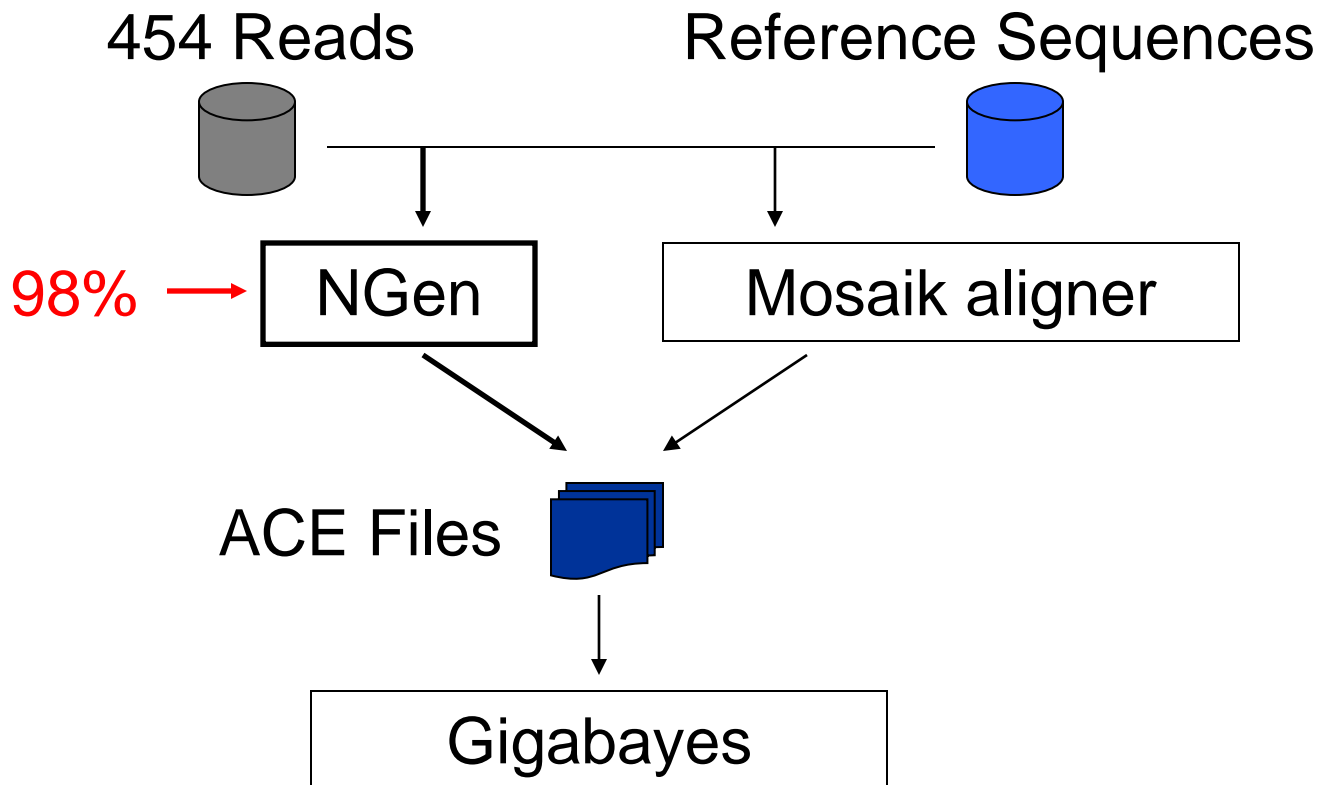
- 3'Capture data for 2 parental lines and 18 Segregation lines
- High Quality reference contigs generated
- **Trialed 3 methods for SNP calling**

ACE + Polybayes

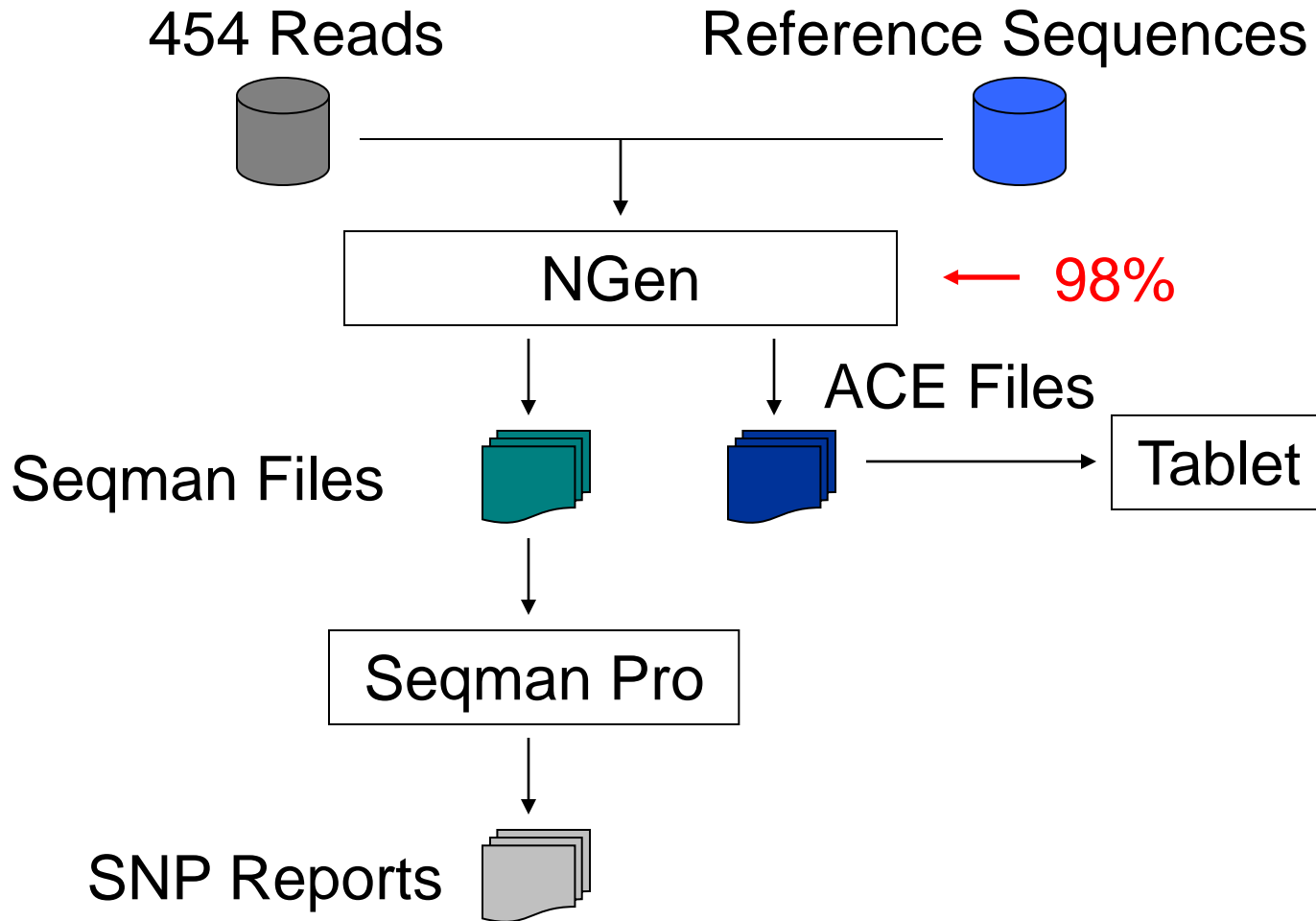


Gigabayes

- Developed by the Marth lab
(<http://bioinformatics.bc.edu/marthlab/GigaBayes>)



DNAStar



Summary of Methods

SNP Calling Method	Running Time	Amount of Manual Intervention Required	Quantity of High-Quality SNPs
Barbazuk et al.	2 Days	High	Moderate
ACE + Polybayes	2 Days – 7 Days	Moderate	Good
DNASTAR	2 Days	Moderate - High	Good
Gigabayes	2 Days – 7 Days	Moderate – High	N/A

Methods Under Investigation

- Gigabayes
- MAQ
- CLC Bio
 - Initial thoughts are positive
 - Will be testing it with both Illumina and 454 data

Downstream analysis of SNP Reports

- **SNP Reports**

- Individual reports for each line/genotype
- Export SNP report (Seqman Pro)
- Parse to compare all

- **Custom Perl Pipeline**

- Track all called SNPs
- Compare in same position as confident SNPs
- Report if same as reference (0)
- Or no sequence data at that position (X)
- Final spreadsheet format

Sample of Lentil scoring matrix

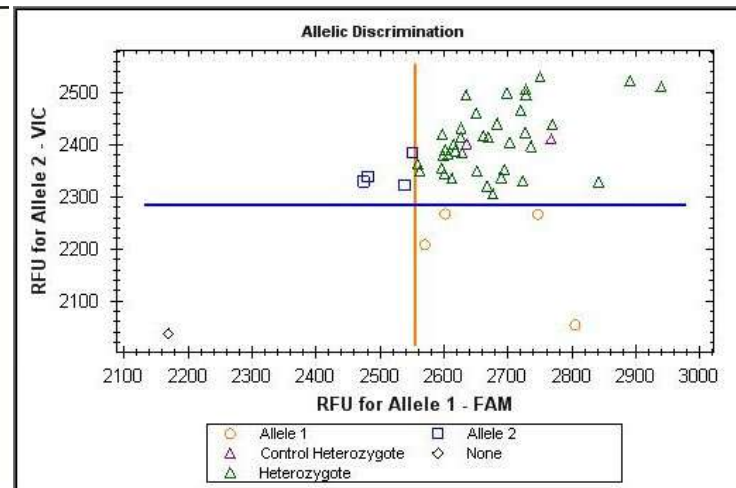
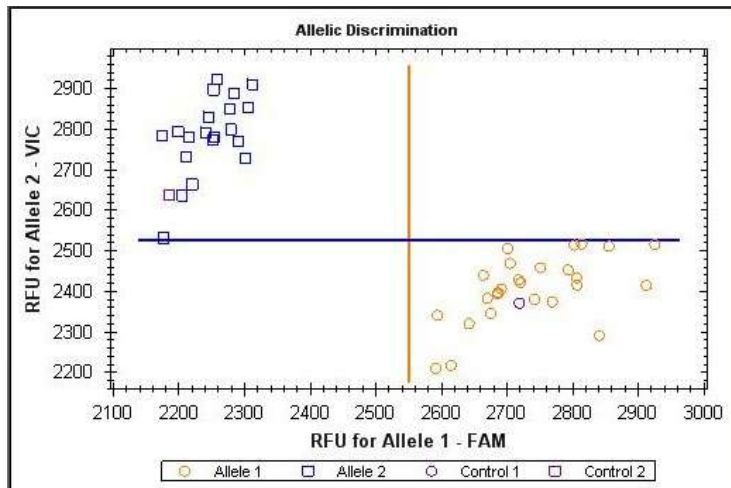
Contig	Reference Pos	Redberry (ref)	964A	CDC_Milestone	Eston-A	ILL-8006	ILL5588
00003	154	C	0	0	A	A	A
00003	492	G	0	0	100	X	A
00004	206	C	0	0	G	G	G
00010	74	A	0	0	0	G	G
00010	76	C	0	0	0	A	A
00010	472	A	0	0	0	G	G
00010	518	C	0	0	0	100	T
00012	343	T	C	C	C	0	C
00012	469	A	T	T	T	0	T
00012	679	T	A	A	A	0	A
00012	738	A	C	C	C	0	C
00012	788	A	C	C	C	0	C
00042	225	G	100	50	66.7	A	100

Sample of Brassica scoring matrix

CONTIG	DH1075 ref	PSA12	SG-10	SG-168	SG-201	SG-221
194_10006_1-373	+	-	+	+	+	+
194_10006_1-499	+	-	+	+	+	+
194_10006_1-518	+	-	+	+	+	+
194_12112_1-150	+	-	+	+	+	-
194_12112_1-184	+	-	+	+	+	-
194_1005_2-156	+	-	-	+	-	+
194_1005_2-444	+	-	-	-	-	+
194_1005_2-480	+	-	-	-	-	+
194_1005_2-483	+	-	-	-	-	+
194_1005_2-525	+	-	-	-	-	+
194_1005_2-667	+	-	X	-	X	+

Validation of SNPs

- Done using the KASPar method



Array Design

- Format data for submission to Illumina for scoring of SNPs
- Annotation of contigs using closely related species
- Selection of SNPs based on Illumina score and even distribution across genome

Future Work

- Move towards shotgun transcriptome and genomic SNP discovery using Illumina RNA-Seq and re-sequencing data
- Further evaluate tools for SNP detection
- Produce reference mapping + SNP calling tools that can integrate 454 and Illumina data if no suitable method can be found

Challenges of SNP Discovery

- Computation resources
 - CPU time
 - RAM usage
- Determination of High Quality SNPs
 - Creation of a high quality reference
 - Dependant on the reference quality
 - Massive number of SNPs to sort through
 - Platform specific problems (i.e. Indels)
 - Extension of sequences past the reference by the assembly algorithm
 - Complicated by ploidy levels
- Making SNP discovery accessible to all researchers

Acknowledgements

- AAFC
 - Isobel Parkin
 - Erin Higgins
 - Matthew Links
 - Lily Tang
 - Steve Robinson
- PBI
 - Andrew Sharpe
 - Christine Sidebottom
 - Larissa Ramsay
 - Kishore Gali
- University of Saskatchewan
 - Kirstin Bett
 - Lacey Sanderson
- Funding
 - The Agriculture and Agri-Food Canada ABIP PURENET project
 - AAFC and NRC-PBI Genomics Health Initiative



Canada